The New Teacher Project

RESETTING RACE TO THE TOP

December 2010

Why the Future of the Competition Depends on Improving the Scoring Process

POLICY BRIEF

POLICY BRIEF | December 2010

Resetting Race to the Top

Why the Future of the Competition Depends on Improving the Scoring Process

What did it take to win Race to the Top?

In a period of less than a year, dozens of states made major policy changes and enlisted a broad range of stakeholders to support ambitious school reform plans that often ran more than 1,000 pages in length. They were pursuing a share of approximately $4.3 billion in federal stimulus funding to turn their ideas into reality during a time of extreme fiscal shortage. The result was more state-level education reform than occurred in the previous two decades combined.

Race to the Top represented a new paradigm in federal education. Instead of spreading relatively modest dollars evenly across all jurisdictions through funding formulas—as virtually all federal education funding has been and continues to be spent—a small number of successful states received all of the available funding, and in turn made it available only to local districts that voluntarily agreed to participate in the state's plan. The approach hinged on a review process that could credibly identify the best applications.

Race to the Top hinged on a review process that could credibly identify the best applications.

The New Teacher Project actively participated in the Race to the Top process. We viewed the competition as an exceptional opportunity to accelerate needed change. Prior to the first application deadline, we published a summary of the contest guidance and a set of policy recommendations for states. We collaborated with many states on their applications and, in April, we analyzed the scores of the first round finalists. At the time, we commended Secretary of Education Arne Duncan for setting a bold vision for reform and holding a high bar for state applicants. But we also identified a series of potentially serious problems in the scoring process.

Specifically, we noted a lack of differentiation in some areas of scoring, general rating inflation, deviation from the scoring guidance, and excessive influence of outlier ratings on the final scores. As we wrote at the time, these problems created the possibility that "[l]ess-deserving states could win at the expense of states truly committed to and capable of dramatic reform."

The second round of Race to the Top winners were named early this fall and, in November, the Department of Education announced that it would conduct a "lessons learned" review of its competitive grant programs, including Race to the Top. Our analysis of the second round scoring process suggests a

number of critical issues that such a review should consider carefully and attempt to rectify. We conclude that many winners submitted ambitious proposals backed up by serious plans for implementation, and that there remains much to celebrate about the contest's impact. However, we also find that the scoring problems of the first round persisted, leading to results that do not fully align with the stated goals or spirit of the program.

Difficulties with the scoring process were most evident in the losses of Colorado and Louisiana, states that have made concrete progress where many other applicants have made only promises.[1] But evidence of scoring problems and inconsistencies stretched across many applications. While some variation is to be expected in any complex, judgment-intensive process, the variation in Race to the Top scoring—rather than the merits of the applications themselves—appears to have placed several states in the winning group and shut others out. Put simply, winning Race to the Top required a solid application—but also the luck of the draw.

Winning Race to the Top required a solid application—but also the luck of the draw.

Our analysis suggests that two key factors yielded outcomes that did not fully match the stated priorities of the contest: (1) the review process and tools allowed reviewers too much freedom to assign or deduct points subjectively; and (2) the Department of Education did not exercise its discretion to modify even the most questionable results (most likely to avoid the appearance of politically-motivated intervention).

It is critical to understand that states were not actually compared to one another in the scoring process. Instead, consistent with standing Department policy, each reviewer scored his or her applications only against the contest rubric, applying standards that often varied according to individual interpretation. Reviewers received feedback but made their own final scoring determinations, and some states drew reviewers who rated applications more harshly or leniently than others. The Department then selected winners based on the precise scores of the review teams rather than treating the scores as advisory.

This analysis is not intended to invalidate the plans or achievements of the Round 2 winners or to belittle the hard work or competence of the Race to the Top reviewers. The winning states invested deeply in their work and should be proud of it. The reviewers took on a great burden and carried it out with dedication and diligence. The problem is in the process. Our goal in exploring these issues is to ensure that competitive funding in general, including Race to the Top, continues to be a viable vehicle for supporting the transformation of our schools. The policy breakthroughs sparked by Race to the Top before a single penny was spent prove the power of this approach. The solution, we will argue, requires an improved scoring process that restores credibility in the contest and inspires states to race with confidence.

[1] *Disclosure: The New Teacher Project (TNTP) provided formal and informal guidance to more than a dozen states on their Race to the Top plans as part of our efforts to advance reforms that align with our organizational mission. Among states receiving our help, some were successful, others were not. We worked with both Louisiana and Colorado, states which are highlighted prominently in this analysis. A large number of states mentioned TNTP in their applications – as many as 18 in the first round. Some states expressed an intention to partner with TNTP upon winning a grant while others merely referenced our research. Not all states that mentioned TNTP as a potential partner contacted us before doing so. In addition, we currently operate projects of some form in six of the ten states that received second round grants: Georgia, Maryland, New York, North Carolina, Rhode Island, and Washington, DC., as well as in finalist states Arizona, California, Colorado, Illinois, Louisiana, New Jersey, and Pennsylvania.*

What Went Right

Before exploring the challenges in the Race to the Top scoring process, it is important to acknowledge the many laudable aspects of the competition. Contest administrators faced an incredibly difficult challenge in creating and executing a high-stakes process under tight timelines. There was no possibility that all observers could be satisfied, no matter the outcome. Reviewers were asked to pore over every detail in wide-ranging applications that often exceeded 1,000 pages and, from all available evidence, they carried out their work conscientiously and diligently. In addition, we applaud the following successes:

- **Design:** The design of the contest itself was remarkable. After decades of stagnation and piecemeal efforts to improve schools, Race to the Top invited states to aim for excellence, not compliance with federal mandates. It successfully leveraged a relatively small amount of federal funding to great effect.

- **Guidance**: The competition offered clear priorities and guidance for applicants. Federal officials heavily emphasized coherence of vision and the ability of states to implement their plans. No section carried as many points as "Great Teachers and Leaders," implicitly acknowledging the well-documented fact that putting the best educators in our schools is critical to improving student outcomes. Administration officials should be commended for resisting significant pressure to water down program requirements.

- **Engagement:** Secretary Duncan treated Race to the Top as more than a grant program—he seized the opportunity to establish a national sense of urgency around education reform, drawing new resources and energy into the effort. Race to the Top effectively reset the education dialogue.

- **Rigor:** Only two winners were selected in Round 1, setting a high bar and resulting in a significant number of strong state reforms between Rounds 1 and 2.

- **Transparency:** The contest was conducted with extraordinary transparency. Full applications were made available for public review, as were finalist scores, video footage of finalist interviews and review comments—all in a timely fashion.

- **Inter-rater reliability:** From Round 1 to Round 2, consistency among reviewers of the same application improved moderately, such that dropping the highest and lowest scores from each panel of five reviewers would not have changed the winning group.

Despite doing almost everything right with Race to the Top, what mattered most was which states won and what was in their applications. Unfortunately, while the process by which the Department reviewed and scored each proposal was earnest and methodical, it was marred by inconsistent standards that resulted in flawed outcomes.

Flaws in the Review Process

There are two main levers for ensuring that the strongest applications prevail in a peer review process. The first is to manage meticulously the work of reviewers, challenging judgments that lack compelling evidence and insisting on a common standard. This approach may require superseding reviews by managers who look across applications for confirmation of consistency. The second is to accept that there will be flaws in any peer review process, but to reserve discretion to treat peer review scores as merely advisory, with final decisions to be made at the executive level.

Our analysis suggests that neither lever was used effectively. Reviewers had broad discretion to interpret standards for scoring. There does not appear to have been any superseding review that effectively ensured that teams applied generally consistent standards in awarding high, medium, or low points to each application component. Each reviewer saw only a few applications; they had no way of knowing what was in other applications or how they were scored.

Letting go of the wheel entailed real risk; in the absence of executive intervention, inaccurate or inconsistent scoring would become determinative.

Secretary Duncan technically retained the authority to overrule review scores, but he chose not to use this post-hoc power. His decision is understandable; a hands-off approach insulates the contest from politics. But letting go of the wheel entailed real risk; in the absence of executive intervention, inaccurate or inconsistent scoring would become determinative. Below, we outline a number of problematic scoring trends that appear to have had this effect.

(Note: This analysis focuses on issues pertaining to teacher effectiveness, The New Teacher Project's area of expertise. However, the evidence suggests that similar scoring problems exist across content areas.)

1. Inflated ratings

As in Round 1, reviewers were extremely generous in their scoring overall, especially among applicants named finalists. On individual items, reviewers almost always assigned points in the "high" range (as defined by contest guidelines) and almost never assigned a score in the "low" range. In fact, *eight states* exceeded the score that earned Tennessee one of just two awards in the first round. While states surely improved upon their applications from Round 1 to Round 2, the large number of states earning exceptional scores suggests that inflation actually increased in second round scoring.

For finalist applicants, reviewers almost always assigned points in the "high" range and almost never assigned a score in the "low" range.

Consider Section D, "Great Teachers and Leaders." This section was the most politically delicate in the competition. States were asked to submit plans that, among other things, tied teacher and school leader evaluations to student results in a "significant" way; used evaluations to inform decisions that have almost never been informed by performance in the past, such as compensation, certification, and tenure; and equalized distribution of the most effective teachers between high- and low-poverty schools. Meeting the standards for this section with a bold yet actionable plan was extremely challenging.

Yet the Race to the Top review process appears to have allowed leniency in the scoring of Section D. On each item, reviewers could award high, medium, or low points. In Round 1, reviewers assigned "high" points to finalists 73 percent of the time for Section D items, and "medium" points in the remainder of cases; no reviewer assigned "low" points to any finalist on any Section D item. In Round 2, this trend became more pronounced. Reviewers assigned high points to 86 percent of Section D items for finalist states, and medium points to the remainder; once again, no state received low points.

The following chart shows how often high and medium points were awarded on Section D in Round 2. Green shading indicates a score in the high range; yellow shading indicates the medium range.

Score Ranges: Race to the Top Section D

Criterion	MA	NY	HI	FL	RI	DC	MD	GA	NC	OH	NJ	AZ	LA	SC	IL	CA	CO	PA	KY
(D)(1)	21	19	14	18	17	20	20	17	18	15	17	15	17	14	18	13	17	10	18
(D)(2)(i)	5	5	5	4	4	5	5	5	5	4	4	5	4	4	3	4	4	3	5
(D)(2)(ii)	15	13	14	12	11	13	15	14	14	15	15	14	12	14	12	14	10	14	10
(D)(2)(iii)	9	9	10	7	10	9	8	9	10	9	9	9	9	9	7	7	8	8	9
(D)(2)(iv)	24	28	26	25	26	27	25	26	23	27	25	25	23	25	25	22	20	20	24
(D)(3)(i)	14	13	14	12	13	12	14	12	12	12	13	12	12	14	12	13	9	10	9
(D)(3)(ii)	9	9	9	8	10	7	9	7	8	8	9	8	8	10	7	8	8	7	6
(D)(4)	13	14	12	12	12	10	12	11	11	12	14	11	12	10	11	13	12	9	13
(D)(5)	17	18	19	18	18	14	19	19	17	19	18	17	14	12	16	15	17	14	18

"High" Score "Medium" Score "Low" Score

But how do we know that applications did not improve so much that they truly warranted generous marks? In part because objective reviews from groups like the National Council on Teacher Quality (NCTQ) suggest otherwise. Prior to the announcement of results, NCTQ reviewed the same materials that were available to RTTT reviewers and placed the Section D plans for all 19 finalists into three categories: red, yellow, and green. Just four states received a green rating. Eight of the 19 finalists received a red rating, reflecting significant deficiencies. Although NCTQ's assessment used criteria that did not precisely mirror those of Race to the Top, the discrepancy between the two sets of ratings is often striking. For example, Massachusetts received high points in every category for Section D, but it received a red rating from NCTQ, which wrote that the state's approach "puts Massachusetts way behind others in committing to details" on educator evaluations.

Section D should have been an opportunity for states like Colorado and Louisiana to set themselves apart from the rest of the field; unlike many other finalists, both states mustered great political will to pass legislation consistent with Race to the Top guidelines. Instead, many states achieved very high scores that belie what we know to be true: that although states are beginning to take laudable steps to improve educator effectiveness, most are starting with a collection of promises that could easily evaporate at any moment. Their plans should have been viewed cautiously. Inflation masked the unique strengths of the truly leading states and made Section D less meaningful than it was intended to be.

2. Inconsistent scoring

In addition to being vulnerable to rating inflation, the Round 2 review process also failed to ensure adequate consistency in scoring. This problem manifested itself in two major ways: First, review teams often seemed to apply different scoring standards to similar application material; and second, application material that had received one score in the first round often received a significantly different score when examined by a new review team in the second round.

The scoring of Illinois' application in Round 2 illustrates both aspects of the problem vividly, especially in Section A1, which focuses on the commitment of school districts and teachers' unions to the state's plan.

In many ways, Illinois represented the best spirit of labor-management collaboration in Race to the Top. The state passed five major pieces of legislation to support its application—all five with the support of its largest teachers' union. The union actively enlisted districts and union locals as participants in the state's plan, and its executive director was one of five state representatives at the in-person interview for the competition's finalists. All told, Illinois secured participation from districts representing 81 percent of students in the state, including 86 percent of students eligible for free and reduced lunch; 49 percent of participating districts also received local union support.

On district commitment, Illinois earned fewer points in Round 2 than in Round 1, despite improving *support from both districts and unions between rounds.*

Yet in the section on Local Education Agency (LEA) commitment, Illinois received 35 out of 45 points in Round 2—four points *fewer* than it received in Round 1, when it had substantially *less* support from both districts and unions. Illinois also saw its score drop in other sections that had scored highly in the first round and that it had reasonably left untouched.[2]

Other states seem to have been judged more gently. California's application, for instance, seemed undeniably weaker than Illinois on LEA commitment. It included districts representing just 28 percent of its students, and 68 percent of its students on free and reduced-price lunch. Just 33 percent of districts were able to secure union support, and the state's largest teachers' union vociferously opposed the entire Race to the Top program. Inexplicably, California received exactly the same average score for LEA commitment as Illinois: 35 points.

Similarly, Ohio's application was joined by districts representing 62 percent of its students and 66 percent of its free and reduced-price lunch students—both totals significantly below Illinois. While Ohio did receive union support in all participating LEAs, this support had not been tested as it had in Illinois, where unions actively supported enshrining major reforms into five state laws. Nonetheless, Ohio received 41 out of 45 points—six points more than Illinois.

[2] It is important to note that variation in scoring from round to round did not affect each state equally. For instance, while Illinois lost at least three full points on four different items, no such thing happened to North Carolina, which dropped no more than one point on any item compared to its Round 1 score. Similarly, New York's biggest drop was 1.2 points. Massachusetts, the highest scoring application in the second round, declined no more than 0.6 points on any item.

Ohio's strong performance may suggest that union participation carried heavy sway. But Maryland's experience tells a different story. District participation in Maryland's plan was comparable to that of Illinois; the state signed on districts representing 79 percent of students and 85 percent of students receiving free and reduced price lunch. But Maryland received support from just 2 of 22 union locals in participating LEAs. Even so, it earned a score of 38 out of 45, or three points higher than Illinois.

It is inevitable that separate review panels will interpret state plans differently to some degree, and reviewers may have based their scores on other factors in addition to those considered here. But states have a right to expect that similar material will achieve relatively similar results. A plan deemed best-in-class by one reviewer should not be interpreted as poor by another. A common definition should guide review teams in assessing factors such as strength of LEA participation. To see such fluctuation so regularly suggests that luck was a large factor in a contest that was meant to reward will.

Why did reviewers rate applications from California, Ohio, and Maryland more leniently than Illinois when it came to LEA participation? What did Illinois lack that the other districts included? How could Illinois score lower in Round 2 even after *improving* its level of participation?

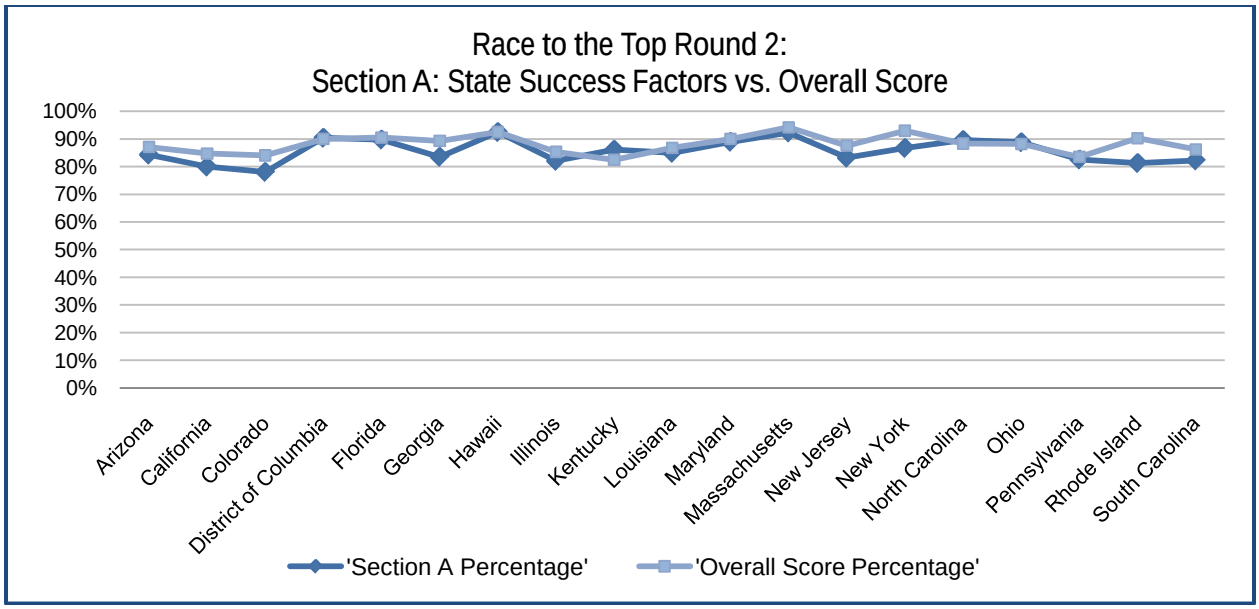
These are all fair questions. And without answers, state education officials and union leaders may also reasonably ask: Why should we take the risk of supporting bold legislative efforts and building partner buy-in when other states that appear to have done less may score just as well—or better?

3. Subjective scoring

Our analysis suggests that some scoring inconsistencies may have been the product of a larger pattern of reviewer subjectivity. In some cases, it appears that reviewers who generally approved of an application tended to rate all parts of it positively, while reviewers with a lower opinion of the application tended to deduct points freely. This trend raises questions about whether the review process allowed subjectivity to trump evidence.

One indication of subjective judgment is the high correlation between scores on separate application sections dealing with unrelated content. For example, in the second round, the ten winning states each earned 88 percent or more of the available points (Ohio had the lowest winning score, with 441 out of 500 points, or 88.2 percent). A strong score in Section A, which concerns state success factors, was an excellent predictor of overall success; all seven of the applications that scored 88 percent or better in Section A won funding. Just three of the other 36 applications—those below 88 percent on Section A—won grants.

The following graph depicts the uncanny similarity between scores on Section A and overall application scores for the Round 2 finalists.



Our analysis suggests that scores for Section A were predictive of scores on other sections, especially Section D ("Great Teachers and Leaders"). The average Section D score for finalists with Section A scores above 88 percent was 121. For all other finalists, the average was just 114.

This trend is surprising given that Sections A and D are so dissimilar. Section A focuses on cohesiveness of state vision, level of buy-in, and past record of academic success. It reflects the big picture. Section D, on the other hand, is very specific. It addresses systems for measuring and improving the performance of educators and prioritizes particular policy directions, such as using evidence of student growth in educator evaluations. There is no reason to assume that states that excelled in Section A would excel in Section D. But when reviewers had a less favorable opinion of Section A, they appear to have been grudging in their view of provisions in Section D—to a worrisome degree.

For example, reviewers for Louisiana, which scored below average for finalists on Section A, did not award 10 out of 10 points in Section D for annual teacher evaluations, despite what appears to be unmistakable evidence that the state deserved the full allotment of points. In fact, Louisiana state law *requires* annual evaluations that result in formal, written feedback within 15 days. A data system will provide access for educators to student data. One of Louisiana's evaluators acknowledged all of these strengths of the state's plan—and then assigned 8 out of 10 points, without explanation.

In contrast, Hawaii, which received the top score in the entire contest on Section A, received 10 points from every single reviewer in the part of Section D that addresses annual evaluations, even though the state currently mandates evaluations for experienced teachers every five years and is only beginning to shift to annual evaluations. The Hawaii review team appears to have applied a lenient standard in its scoring after developing a favorable impression of the state's application overall.

While two points on one item may not seem meaningful, the loss or gain of a few points here or there made a difference. In the end, Hawaii scored 122 points in the "Great Teachers and Leaders" section, compared to just 111 for Louisiana, which was viewed by many (including TNTP) as having one of the

best human capital plans in the entire contest. In fact, Louisiana's application had already received the top Great Teachers and Leaders score—*in Round 1.* Simply equaling Hawaii's score in this area would have bumped Louisiana from 13th place to 9th, more than enough to gain an award. Had Louisiana received the best Great Teachers and Leaders score in the contest (128 out of 138, by New York), it would have placed 5th and won a grant. This single section had the power to change Louisiana's outcome.

In summary, it is unclear whether reviewers truly scored each section independently and based on evidence or formed an overall judgment that—perhaps unconsciously—influenced their scoring across the application. It is clear, however, that there are multiple instances where state scores show a general trend of lenient or harsh scoring that is not well supported by reviewer comments, begging the question of whether subjectivity played an outsized role.

4. No accounting for depth of commitment

While preparing their Race to the Top proposals, some states channeled massive political will to enshrine reforms in state law. Dialogue in these states was not always easy, and not every participant left happy with the outcome. But states in this situation were able to present an unambiguous, complete commitment; their plans carried the force of law.

Strangely, states that made the strongest and most durable commitments did not always win the most points. Meanwhile, states that adopted reforms more tentatively, through carefully worded memoranda of understanding or vague declarations of intent, scored better despite leaving many of the thorniest questions unresolved. This trend was especially disappointing because it was firmly at odds with the contest's goals and rhetoric.

Colorado, for example, passed a major piece of legislation, Senate Bill 191, which aligned closely to the requirements for Section D2 of Race to the Top. State law in Colorado now requires teacher evaluations to be based predominantly on evidence of student performance, which goes beyond the contest requirement of "significant" emphasis on student performance. But most importantly, state law describes how evaluations must be used to inform teacher development, promotion, non-probationary status, and dismissal—the main categories addressed by the competition's guidance. Even so, Colorado earned only modest points. The average score among five reviewers for Section D2(iv), the section pertaining to these issues, was 20 out of 28.

Hawaii, in contrast, relied in Section D2(iv) on an "Agreement of Concepts" signed by its lone school district and teachers' union. The agreement included aspirational statements such as "Conceptually, the parties agree that a minimum continuous three year probationary period prior to earning tenure is ideal." On teacher compensation, Hawaii outlined an intention to create a system of performance-based pay at a future date, offering that while "details still need to be officially negotiated, informal discussions" with the teachers' union had occurred. On the issue of dismissing ineffective teachers, Hawaii made no changes to existing policy, though it planned additional training for administrators. The following table summarizes the differences between Hawaii's and Colorado's plans.

Section D2(iv) Components	Hawaii's Plan	Colorado's Plan
Developing teachers and principals	Principals and teachers will be required to create development plans that are updated at least every two years for effective teachers, more often for teachers with lower ratings.	Teachers and principals will have individual development plans informed by previous year's evaluation. Plans must include professional development goals.
Compensating promoting, retaining	The state is currently negotiating a new teacher compensation system with the teachers' union. The application envisions the system being similar to Denver's ProComp approach and expresses an "intent" for the new system to incorporate student growth.	Under state law, all LEAs in the state will use teacher evaluations to inform compensation. Participating LEAs must use evaluation ratings to inform compensation by 2013, including additional compensation for highly effective teachers.
Granting tenure and/or certification	The state and union have pledged to include a new tenure process—which would require three years of effective ratings in a five year period—in the next union contract, which has yet to be negotiated. The state teacher standards board has also agreed to "consider" using evaluations in re-licensing process.	By state law, all LEAs, participating and not participating, will use teacher evaluations in tenure decisions. Participating LEAs will do it by 2013. Three ratings of effective or better will earn tenure, and two ratings of ineffective will result in the loss of tenure.
Removing ineffective educators	The state's collective bargaining agreement says that ineffective teachers "shall be terminated." The application acknowledges that while this "authority has not been widely used in the past, the current Superintendent is committed to using it when necessary and appropriate." The application further states that "some changes to tenure and termination procedures for both teachers and principals likely will need to be re-examined through the collective bargaining process."	Under state law, all LEAs will use teacher effectiveness ratings to inform decisions about dismissal. Additionally, state law requires that all staffing reductions, including building-level reductions and system-wide layoffs, will be executed with effectiveness as the first criterion considered.
Total points awarded	26	20

On almost every dimension, Hawaii's proposal for Section D2(iv) appears to be weaker than Colorado's. But during the review process, it not only equaled the Colorado score—it substantially *exceeded* it. Hawaii was awarded an average of 26 out of 28 points, compared to Colorado's 20.

Despite the national praise given to Colorado's SB 191 by teacher policy experts, Hawaii was not alone in outscoring Colorado in this area. An astonishing *18 states* surpassed Colorado's score of 20. This may have been understandable if Colorado had put forward an application that simply did not represent its strengths in this area as compellingly as other states, but in fact Colorado provided a detailed summary of SB 191 in its proposal. The problem appears to be in how the evidence that states presented in support of their plans was interpreted and valued.

These scoring outcomes give the impression that states with good intentions to improve policies could be viewed on equal footing with states that had already exercised the political courage to change them prior to applying for a grant. In fact, they suggest that offering less detail could actually be an advantage. States might reasonably ask: Why go through the arduous process of weaving policies into the fabric of state law when you can merely agree to consider them?

The Consequences

Taken together, the scoring trends described above produced a cluster of applications with scores above 400 out of 500 possible points. Our review suggests that the relative positioning of the finalists is difficult to justify. Despite legitimate strengths, the winning group seems to have benefitted significantly from the vagaries of the scoring process. Had the same applications been scored again by new review teams, it is not inconceivable that half or more of the winning group could have been changed.

It is undeniable that some of the proposals that best embodied the spirit and substance of Race to the Top were not winners.

There is much to applaud in each state's application, and there are sure to be breakthrough accomplishments among the winning states as they implement their proposals. But it is undeniable that some of the proposals that best embodied the spirit and substance of Race to the Top—and which seemed to align most closely to the contest rubric—were not winners. Several were not even close.

The failure of states like Colorado, Louisiana, and Illinois to win grants based on the scoring trends described in this analysis creates a worrisome distance between the rhetoric of Race to the Top and the contest results. While representatives of the Department of Education, including Secretary Duncan, repeatedly declared that "watered-down proposals with lots of consensus won't win… proposals that drive real reform will win," the variation from review team to review team reflects that the process did not always live up to the Secretary's mandate. Department leaders did not put themselves in a position to ensure that only "real reform" proposals won grants. The decisions were actually in the hands of dozens of separate review teams, interpreting guidance as best they could, without a broad perspective of what was in the applications not assigned to them. Sometimes, "real reform" won; sometimes it did not.

Moving Forward

We applaud the spirit of innovation and possibility that Race to the Top unleashed and the incredible effort invested in it by education leaders and policymakers, Department of Education staff, and many others. It has been a generational accomplishment in educational reform that demonstrates the power of carefully structured competitive funding programs to jumpstart change and rally diverse stakeholders around a common goal. We continue to believe that competitive funding must be meaningful part of federal education strategy.

Still, despite everything that Secretary Duncan and the Department of Education did right with this program, the results may undermine confidence in the process, especially among states that took big risks to align their policies with the competition's priorities, only to be shut out. If Race to the Top is continued, several critical changes can improve the contest and make it a stronger vehicle for long-term reform. Most especially, the Department of Education must more actively manage the work of reviewers, enforcing consistent scoring from one team to another through cross-application review processes. These changes were necessary after the first round; they are now overdue.

Given the erratic scoring outcomes in both rounds of the competition, it is also time to consider executive discretion in the process. The peer review process should still determine which states become finalists and should play a significant role in setting awards. But especially in the absence of stricter scoring controls, peer scores cannot be treated as precise, unquestionable determinations. The Secretary should be prepared to make final decisions based on a combination of the peer scores and any contextual factors, with the goal of ensuring that the contest's outcomes match its priorities.

Failure to improve the scoring process is likely to reduce the quality and impact of any future contest rounds. States will have a difficult time making the case for ambitious reforms if they cannot be confident that their risks will be rewarded. For Race to the Top truly to deliver on its promise, it must be improved.

About The New Teacher Project

The New Teacher Project (TNTP) strives to end the injustice of educational inequality by providing excellent teachers to the students who need them most and by advancing policies and practices that ensure effective teaching in every classroom. A national nonprofit organization founded by teachers, TNTP is driven by the knowledge that effective teachers have a greater impact on student achievement than any other school factor. In response, TNTP develops customized programs and policy interventions that enable education leaders to find, develop and keep great teachers. Since its inception in 1997, TNTP has recruited or trained approximately 43,000 teachers—mainly through its highly selective Teaching Fellows™ programs—benefiting an estimated 7 million students. TNTP has also released a series of acclaimed studies of the policies and practices that affect the quality of the nation's teacher workforce, including *The Widget Effect* (2009) and *Teacher Evaluation 2.0* (2010). Today TNTP is active in more than 25 cities, including 10 of the nation's 15 largest.

www.tntp.org